programa final

Encontro Nacional de Produtores e Usuários de Informações Sociais, Econômicas e Territoriais

INFORMAÇÃO PARA UMA SOCIEDADE MAIS JUSTA

III Conferência Nacional de Geografia e Cartografia

IV Conferência Nacional de Estatística

Reunião de Instituições Produtoras
Fórum de Usuários
Seminário "Desafios para Repensar o Trabalho"
Simpósio de Inovações
Jornada de Cursos
Mostra de Tecnologias de Informação

27 a 31 de maio de 1996
Rio de Janeiro, RJ BRASIL

Uma das maneiras de olhar o ofício de produzir informações sociais, econômicas e territoriais é como arte de descrever o mundo. Estatísticas e mapas transportam os fenômenos da realidade para escalas apropriadas à perspectiva de nossa visão humana e nos permitem pensar e agir à distância, construindo avenidas de mão dupla que juntam o mundo e suas imagens. Maior o poder de síntese dessas representações, combinando, com precisão, elementos dispersos e heterogêneos do cotidiano, maior o nosso conhecimento e a nossa capacidade de compreender e transformar a realidade.

Visto como arte, o ofício de produzir essas informações reflete a cultura de um País e de sua época, como essa cultura vê o mundo e o torna visível, redefinindo o que vê e o que há para se ver.

No cenário de contínua inovação tecnológica e mudança de culturas da sociedade contemporânea, as novas tecnologias de informação - reunindo computadores, telecomunicações e redes de informação - aceleram aquele movimento de mobilização do mundo real. Aumenta a velocidade da acumulação de informação e são ampliados seus requisitos de atualização, formato - mais flexível, personalizado e interativo - e, principalmente, de acessibilidade. A plataforma digital vem se consolidando como o meio mais simples, barato e poderoso para tratar a informação, tornando possíveis novos produtos e serviços e conquistando novos usuários.

Acreditamos ser o ambiente de conversa e controvérsia e de troca entre as diferentes disciplinas, nas mesas redondas e sessões temáticas das Conferências Nacionais de Geografia, Cartografia e Estatística e do Simpósio de Inovações, aquele que melhor enseja o aprimoramento do consenso sobre os fenômenos a serem mensurados para retratar a sociedade, a economia e o território nacional e sobre as prioridades e formatos das informações necessárias para o fortalecimento da cidadania, a definição de políticas públicas e a gestão político - administrativa do País, e para criar uma sociedade mais justa.

Simon Schwartzman
Coordenador Geral do ENCONTRO

Promoção

Fundação Instituto Brasileiro de Geografia e Estatística
IBGE
Fundação Instituto Brasileiro de Geografia e Estatistica
IBGE
Associação Brasileira de Estudos Populacionais
ABEP

Co-Promoção

Associação Brasileira de Estatística
ABE
Associação Brasileira de Estudos do Trabalho
ABET
Associação Brasileira de Pós-graduação em Saúde Coletiva
ABRASCO
Associação Nacional de Centros de Pós-graduação em Economia
ANPEC
Associação Nacional de Pós-graduação e Pesquisa em Ciências Sociais
ANPOCS
Associação Nacional de Pós-graduação e Pesquisa em Geografia
ANPEGE
Associação Nacional de Pós-graduação e Pesquisa em Planejamento Urbano e Regional
ANPUR
Sociedade Brasileira de Cartografia
SBC

Apoio

Federação das Indústrias do Estado do Rio de Janeiro
FIRJAN
Academia Brasileira de Letras
ABL
Conselho Nacional de Pesquisas
CNPq
Financiadora de Estudos e Projetos
FINEP
Revista Ciência Hoje

Institutos Regionais Associados

Companhia do Desenvolvimento do Planalto Central
CODEPLAN (DF)
Empresa Metropolitana de Planejamento da Grande São Paulo S/A
EMPLASA (SP)
Empresa Municipal de Informática e Planejamento S/A
IPLANRIO (RJ)
Fundação Centro de Informações e Dados do Rio de Janeiro
CIDE (RJ)
Fundação de Economia e Estatística
FEE (RS)
Fundação de Planejamento Metropolitano e Regional
METROPLAN (RS)
Fundação Instituto de Planejamento do Ceará
IPLANCE (CE)
Fundação João Pinheiro
FJP (MG)
Fundação Joaquim Nabuco
FUNDAJ (PE)
Fundação Sistema Estadual de Análise de Dados
SEADE (SP)
Instituto Ambiental do Paraná
IAP (PR)
Instituto de Geociências Aplicadas
IGA (MG)
Instituto de Pesquisas Econômicas, Administrativas e Contábeis
IPEAD (MG)
Instituto do Desenvolvimento Econômico Social do Pará
IDESP (PA)
Instituto Geográfico e Cartográfico
IGC (SP)
Instituto de Apoio à Pesquisa e ao Desenvolvimento "Jones dos Santos Neves"
IJSN (ES)
Instituto Paranaense de Desenvolvimento Econômico e Social
IPARDES (PR)
Processamento de Dados do Município de Belo Horizonte S/A
PRODABEL (MG)
Superintendência de Estudos Econômicos e Sociais da Bahia
SEI (BA)

Coordenação Geral

Simon Schwartzman

Comissões de Programa

Confege

César Ajara (IBGE)
Denizar Blitzkow (USP)
Jorge Marques (UFRJ)
Lia Osório Machado (UFRJ)
Mauro Pereira de Mello (IBGE)
Speridião Faissol (UERJ)
Trento Natali Filho (IBGE)

Confest

José A. M. de Carvalho (UFMG)
José Márcio Camargo (PUC)
Lenildo Fernandes Silva (IBGE)
Teresa Cristina N. Araújo (IBGE)
Vilmar Faria (CEBRAP)
Wilton Bussab (FGV)

Comissão Organizadora

Secretaria Executiva - Luisa Maria La Croix
Secretaria Geral - Luciana Kanham
Confege, Confest e Simpósio de Inovações
Anna Lucia Barreto de Freitas, Evangelina X.G. de Oliveira, Jaime Franklin Vidal Araújo, Lilibeth Cardozo R.Ferreira e Maria Letícia Duarte Warner
Jornada de Cursos - Carmen Feijó
Finanças - Marise Maria Ferreira
Comunicação Social - Micheline Christophe e Carlos Vieira
Programação Visual - Aldo Victorio Filho e Luiz Gonzaga C. dos Santos
Infra-Estrutura - Maria Helena Neves Pereira de Souza
Atendimento aos Participantes - Cristina Lins
Apoio
Andrea de Carvalho F. Rodrigues, Carlos Alberto dos Santos, Delfim Teixeira, Evilmerodac D. da Silva, Gilberto Scheid, Héctor O. Pravaz, Ivan P. Jordão Junior, José Augusto dos Santos, Julio da Silva, Katia V. Cavalcanti, Lecy Delfim, Maria Helena de M. Castro, Regina T. Fonseca, Rita de Cassia Ataualpa Silva e Taisa Sawczuk

Registramos ainda a colaboração de técnicos das diferentes áreas do IBGE, com seu trabalho, críticas e sugestões para a consolidação do projeto do ENCONTRO.

Análise estatística de dados longitudinais

Julio da Motta Singer

Departamento de Estatística
Instituto de Matemática e Estatística
Universidade de São Paulo

Resumo

Apresentamos uma breve caracterização de estudos com dados longitudinais salientando as diferenças relativamente a outros tipos de planejamentos. Indicamos os principais modelos estatísticos utilizados na análise desse tipo de dados, mostrando que o esforço analítico adicional está concentrado na modelagem da estrutura de correlação entre as observações realizadas na mesma unidade amostral. Discorremos sobre os métodos de estimação e implementação computacional e finalmente ilustramos esses tópicos com um exemplo numérico.

Palavras-chave: Análise de perfis, Correlação intra-unidades amostrais, Curvas de crescimento, Dados longitudinais, Medidas repetidas, Planejamento do tipo painel.

1. Introdução

Em muitas situações, há interesse em se estudar o comportamento de uma ou mais características (aqui chamadas de variáveis respostas) dos elementos de uma ou mais populações ao longo de uma certa escala ordenada. Um exemplo envolveria a comparação da evolução do faturamento de diferentes tipos de empresas ao longo de um certo período de tempo. Num outro campo, um estudo em que o objetivo é avaliar os níveis de poluição em regiões situadas a distâncias crescentes de diferentes tipos de fontes poluidoras constitui outro exemplo. Para efeito de simplificação, referir-nos-emos a essa dimensão ao longo da qual são realizadas as observações, genericamente como **tempo**.

Nesse contexto, em geral, há duas estratégias de coleta de dados: a primeira envolve a observação das variáveis respostas para uma amostra (de empresas, por exemplo) de cada uma das populações em cada instante (e somente nesse instante) ao longo do tempo. A segunda envolve a observação dessas variáveis respostas para os elementos de uma mesma amostra de cada uma das populações durante dois ou mais instantes do período em questão. No primeiro caso dizemos que o estudo é **tranversal** ("cross-sectional") e no segundo caso, dizemos que o estudo é **longitudinal**. Nas áreas de Administração, Economia ou Sociologia, essa forma de coleta de dados também é conhecida como **painel**. Neste ponto, vale ressaltar que o tipo de problema considerado sob a denominação de análise de dados longitudinais difere daquele usualmente tratado na literatura estatística por **análise de séries temporais** pelo fato de que, neste caso, em geral, dispomos de uma única unidade amostral com muitas observações ao longo do tempo (e.g. 100 ou mais), ao passo que naquele, lidamos com várias unidades amostrais (e.g. 5 ou mais) observadas em poucos instantes (e.g. 2 a 10). Vale também lembrar que os planejamentos longitudinais podem ser inseridos

na classe mais ampla dos chamados planejamentos com **medidas repetidas** ("repeated measures"), que incluem, entre outros, os planejamentos do tipo "**split-plot**" e planejamentos com **intercâmbio** ("crossover"). Mais detalhes sobre a conceituação de estudos longitudinais e sua relação com outras estruturas de coleta de dados podem ser encontrados em Goldstein (1979), Singer e Andrade (1986), Duncan and Kalton (1987) ou Diggle, Liang and Zeger (1994), por exemplo.

Estudos longitudinais são de particular interesse quando o objetivo é avaliar **variações** globais ou individuais ao longo do tempo. Em primeiro lugar, esse tipo de planejamento permite observação da(s) variável(eis) resposta(s) sob condições de exposição uniforme das unidades amostrais relativamente a diferentes covariáveis. No exemplo acima, as mudanças no faturamento ao longo do tempo consideradas num estudo longitudinal estariam (pelo menos parcialmente) dissociadas de possíveis diferenças nas políticas administrativas das diversas empresas selecionadas em cada instante de observação de um estudo transversal. Essa característica tem especial interesse quando a variabilidade **entre unidades amostrais** (entre empresas, por exemplo) é maior que a variabilidade **intra-unidades amostrais** (dentro da mesma empresa ao longo do tempo, por exemplo). Em segundo lugar, estudos longitudinais dão subsídios para a avaliação de padrões individuais de variação dos níveis da(s) variável(eis) resposta(s). Finalmente, alguns parâmetros de interesse podem ser estimados de forma mais eficiente sob planejamentos longitudinais do que sob planejamentos transversais com o mesmo número de observações. Essencialmente, essa diferença na eficiência dos estimadores pode ser explicada através do seguinte exemplo.

Consideremos uma situação em que o interesse é comparar as médias de uma variável resposta observada antes e após uma certa intervenção. Denotemos por X a variável observada antes da intervenção e por Y a mesma variável observada após a intervenção. Num planejamento transversal, essa comparação seria realizada a partir dos dados provenientes de duas amostras independentes, cada uma com n unidades amostrais diferentes, digamos $(X_1,\ldots,X_n)$ e $(Y_1,\ldots,Y_n)$ através da estatística $t=(\bar{X}-\bar{Y})/s\sqrt{2/n}$ onde $\bar{X}$ e $\bar{Y}$ são as médias amostrais de X e Y respectivamente e s^2 é uma estimativa da variância de $X-Y$, ou seja, de $\sigma_X^2+\sigma_Y^2$ onde σ_X^2 e σ_Y^2 são respectivamente as variâncias de X e Y. Num planejamento longitudinal, a comparação seria realizada com a observação das variáveis X e Y nas mesmas n unidades amostrais através da utilização da chamada estatística t pareada, dada por $t_d=(\bar{X}-\bar{Y})/s_d\sqrt{1/n}$ onde s_d^2 é uma estimativa da variância de $X-Y$ que neste caso é dada por $\sigma_X^2+\sigma_Y^2-2\sigma_{XY}$ com σ_{XY} denotando a covariância entre X e Y.Convém lembrar que no caso anterior, $\sigma_{XY}=0$, em função da independência estatística entre X e Y.Quando σ_{XY} é positiva, espera-se que o denominador de t_d seja menor que o de t e consequentemente que o teste tenha mais poder para detectar diferenças entre as médias da variável resposta antes e após a intervenção.

As duas principais desvantagens associadas a estudos longitudinais estão relacionadas com seu custo (pois, em geral, é difícil garantir que as unidades amostrais selecionadas sejam observadas nas épocas designadas) e com as dificuldades técnicas de análise dos dados (como veremos adiante, os modelos estatísticos adequados para esse tipo de planejamento são, em geral, mais complexos do que aqueles adequados para estudos transversais).

Essencialmente, os problemas em que estamos interessados quando consideramos estudos longitudinais são os mesmos com que nos deparamos em estudos tranversais e

podem ser classificados do ponto de vista estatístico como problemas de análise de variância (ANOVA) ou, mais geralmente, de regressão (linear ou não-linear). A diferença entre eles, está no fato de que no caso transversal lidamos com dados (estatisticamente) independentes, ao passo que no caso longitudinal é necessário considerar uma possível dependência (estatística) entre eles. O esforço adicional empregado na análise de dados longitudinais relativamente à análise de dados transversais está praticamente concentrado na modelagem da estrutura dessa dependência.

O objetivo da próxima seção é apresentar alguns modelos que permitem incorporar possíveis dependências entre as observações. Concentramo-nos em situações com uma única variável resposta contínua com distribuição gaussiana; modelos para variáveis respostas discretas ou categorizadas podem ser encontrados em Diggle, Liang and Zeger (1994) ou Koch, Singer and Stokes (1992), por exemplo.

2. Principais modelos

Num estudo longitudinal, os dados correpondentes a cada unidade amostral podem ser expressos através de um vetor contendo as respostas em cada um dos instantes de observação e de uma matriz contendo os valores das variáveis explicativas. De uma forma simbólica, o vetor com as p_i respostas associadas à i-ésima unidade amostral (chamado de **perfil de respostas**) pode ser escrito como

$$\mathbf{y}_i = (y_{i1},\ldots,y_{ip_i})^t$$

onde y_{ik} denota a resposta obtida no k-ésimo instante, $(k = 1,\ldots,p_i)$ e $\mathbf{a}^t$ simboliza o vetor $\mathbf{a}$ transposto. Os modelos usualmente empregados na análise têm a forma

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \qquad (2.1)$$

onde $\mathbf{X}_i$ é uma matriz de especificação, cujas colunas contêm os valores das variáveis explicativas (em geral essas variáveis estão associadas à potências do tempo ou são variáveis indicadoras, embora outros tipos de covariáveis também possam ser consideradas), $\boldsymbol{\beta}$ é o vetor de parâmetros que desejamos estimar e $\boldsymbol{\varepsilon}_i$ é um vetor de erros aleatórios com vetor de médias $\mathbf{0}$ e matriz de covariância Σ_i e para o qual geralmente se admite uma distribuição gaussiana.

Num caso em que o objetivo é modelar os dados através de uma reta para explicar a variação das p_i respostas da i-ésima unidade amostral ao longo do tempo, teríamos

$$\mathbf{X}_i = \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_{p_i} \end{pmatrix} \qquad (2.2)$$

onde t_k representa o tempo (em dias, meses etc.) decorrido entre o início do estudo e o instante em que foi realizada a k-ésima observação, e

$$\boldsymbol{\beta} = (a,b)^t$$

onde a e b representam respectivamente o coeficiente linear e o coeficiente angular da reta que pretendemos ajustar. Segundo esse modelo, o valor esperado para uma observação realizada no k-ésimo instante é dado por $a + bt_k$.

Vale a pena notar que os elementos do modelo apresentado acima são os mesmos que utilizaríamos para representar a reta caso o estudo fosse tranversal. O componente que diferencia os modelos para estudos longitudinais é a estrutura imposta à matriz de

covariância intra-unidades amostrais, Σ_i. No caso transversal temos $\Sigma_i = \sigma^2 \mathbf{I}_{p_i}$ onde σ^2 representa a variância comum a todas as observações e $\mathbf{I}_{p_i}$ denota uma matriz identidade de dimensão p_i. Nesse caso a covariância (ou correlação) entre quaisquer duas observações (mesmo que realizadas na mesma unidade amostral) é nula. No caso longitudinal há várias alternativas para a estrutura de Σ_i, tanto mais sofisticadas quanto mais complexa a relação de dependência entre as observações intra-unidades amostrais. A mais simples (conhecida na literatura estatística como **estrutura uniforme**) é aquela em que todas as variâncias são iguais e todas as covariâncias são iguais, ou seja

$$\Sigma_i = \sigma^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} \tag{2.3}$$

onde ρ é o coeficiente de correlação entre quaisquer duas observações realizadas na mesma unidade amostral. A estrutura de covariância mais complexa é aquela em que todas as variâncias podem ser diferentes entre si e todas as covariâncias podem ser diferentes entre si, ou seja

$$\Sigma_i = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p_i} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2p_i} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1p_i} & \sigma_{2p_i} & \cdots & \sigma_{p_i}^2 \end{pmatrix} \tag{2.4}$$

onde σ_k^2 representa a variância das observações realizadas no k-ésimo instante e σ_{kl} representa a covariância entre as observações realizadas na mesma unidade amostral nos instantes k e l. Nesse caso dizemos que a matriz de covariância é **não estruturada**. Se por um lado, a estrutura uniforme tem o atrativo da simplicidade, por outro, ela peca por não permitir a incorporaração de um padrão bastante comum em dados longitudinais, onde as variâncias crescem com o tempo e as correlações decrescem com o espaçamento entre as observações intra-unidades amostrais (ver Kenward (1987), por exemplo). Embora o modelo com matriz de covariância não estruturada admita esse padrão, ele tem a desvantagem de envolver um número muito grande de parâmetros, mesmo nos casos em que o número de medidas intra-unidades amostrais é da ordem de três ou quatro. Esse fato causa problemas tanto de estimação quanto de interpretação. É para resolver esse dilema, que muitos autores têm concentrado esforços para apresentar modelos intermediários que incorporem o padrão de dependência mencionado acima com um número reduzido de parâmetros. Nessa categoria, os modelos mais comuns são os modelos baseados em **processos auto-regressivos** (ver Rao (1967) ou Rochon and Helms (1989), por exemplo), onde

$$\Sigma_i = \sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{p_i} \\ \rho & 1 & \rho & \cdots & \rho^{p_i-1} \\ \rho^2 & \rho & 1 & \cdots & \rho^{p_i-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{p_i} & \rho^{p_i-1} & \rho^{p_i-2} & \cdots & 1 \end{pmatrix} \tag{2.5}$$

ou os chamados **modelos de efeitos aleatórios** (ver Laird and Ware (1982), por exemplo), para os quais a matriz Σ_i não admite uma expressão de interpretação tão simples como

aquelas apresentadas acima, mas também pode ser escrita em função de um número reduzido de parâmetros. Mais especificamente, no caso de ajuste de retas como no exemplo mencionado, um possível modelo de efeitos aleatórios é dado por

$$\Sigma_i = \mathbf{X}_i \Delta \mathbf{X}_i^t + \sigma^2 \mathbf{I}_{p_i} \quad (2.6)$$

onde $\mathbf{X}_i$ é dada por (2.2) e

$$\Delta = \begin{pmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{pmatrix} \quad (2.7)$$

com σ_a^2 e σ_b^2 denotando respectivamente as variâncias dos coeficientes lineares e angulares das retas associadas às diferentes unidades amostrais (consideradas como efeitos aleatórios) e σ_{ab} denotando a correspondente covariância entre esses coeficientes.

De uma forma geral as classes de modelos para a estrutura de covariância envolvem a expressão das matrizes Σ_i como função de um vetor de parâmetros desconhecidos, Θ. No caso do modelo de efeitos aleatórios descrito acima, os parâmetros da estrutura de covariância intra-unidades amostrais, Σ_i correspondem aos quatro elementos do vetor $\Theta = (\sigma_a^2, \sigma_b^2, \sigma_{ab}, \sigma^2)^t$, independentemente do número de observações realizadas em cada unidade amostral.

3. Estimação e implementação computacional

Os métodos de estimação e obtenção de estatísticas de teste sob os modelos utilizados para a análise de dados longitudinais não diferem, substancialmente, daqueles considerados no caso transversal. A maneira mais simples de se analizarem dados longitudinais é através das chamadas **medidas resumo** ("summary measures"); nesse caso, o perfil de respostas de cada unidade amostral é substituido por um único valor (a medida resumo) e o conjunto dessas medidas é analizado transversalmente, em geral através de ANOVA. Exemplos de medidas resumo são a resposta média num determinado período, a resposta máxima, a taxa de variação da resposta nesse período etc. O leitor poderá consultar Domenech (1989) para maiores detalhes. Esse tipo de análise só tem interesse, entretanto, quando os objetivos estão centrados na comparação as populações investigadas com relação a certas características das distribuições da variável resposta e não incluem a avaliação do comportamento da variável resposta ao longo do tempo. Nesse caso, outros métodos de análise, cuja essência descrevemos abaixo, são necessários.

Basicamente, se a estrutura de covariância $\Sigma_i = \Sigma_i(\Theta)$ fosse conhecida, o estimador de mínimos quadrados generalizados (ou de máxima verossimilhança) do vetor β corresponderia a

$$\hat{\beta} = (\sum_{i=1}^{n} \mathbf{X}_i^t \Sigma_i^{-1} \mathbf{X}_i)^{-1} (\sum_{i=1}^{n} \mathbf{X}_i^t \Sigma_i^{-1} \mathbf{y}_i) \quad (3.1)$$

Mesmo em algumas situações onde $\Sigma_i = \Sigma_i(\Theta)$ não é conhecida, é possível obter expressões explícitas para o estimador do vetor β a partir de (3.1). Esse é o caso de situações em que além de o planejamento prever que todas as unidades amostrais sejam observadas nos mesmos instantes, não existam observações incompletas e que as matrizes de covariância sejam do tipo uniforme ou não estruturada. Diz-se, então, que os dados são **balanceados em relação ao tempo**. Estão nessa categoria os problemas tratados na literatura como **Análise de Perfis** ("Profile analysis") e **Análise de Curvas de**

Crescimento ("Growth Curve Analysis"). Maiores detalhes podem ser obtidos em Singer e Andrade (1986) ou Crowder and Hand (1990), por exemplo. Esses estimadores com forma explícita também podem ser obtidos para alguns modelos de efeitos aleatórios quando os dados são balanceados em relação ao tempo. O leitor poderá consultar Graybill (1976) para maiores detalhes. Nesse contexto, a distribuição (exata) do estimador $\tilde{\beta}$. é Normal multivariada com vetor de médias β e matriz de covariância $(\sum_{i=1}^{n} \mathbf{X}_i^t \Sigma_i^{-1} \mathbf{X}_i)^{-1}$.

No caso geral, em que as observações podem ser feitas em instantes não especificados, podem ser incompletas ou podem ter estrutura de covariância regida por outros modelos, temos que recorrer a métodos iterativos para obtenção dos estimadores de máxima verossimilhança ou máxima verossimilhança restrita (que produz estimadores menos viesados). Entre eles, destacamos o método de Newton-Raphson, o método "Scoring" de Fisher ou o método EM. Em particular, o método de Newton-Raphson envolve a especificação de valores iniciais $\beta^{(0)}$ e $\Theta^{(0)}$ para os parâmetros do modelo e o cálculo iterado (para $k = 1,2,\ldots$) da expressão

$$\begin{pmatrix} \beta^{(k)} \\ \Theta^{(k)} \end{pmatrix} = \begin{pmatrix} \beta^{(k-1)} \\ \Theta^{(k-1)} \end{pmatrix} + \begin{pmatrix} \mathbf{J}_{\beta\beta} & \mathbf{J}_{\beta\Theta} \\ \mathbf{J}_{\Theta\beta} & \mathbf{J}_{\Theta\Theta} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{u}_\beta \\ \mathbf{u}_\Theta \end{pmatrix} \qquad (3.2)$$

onde $\mathbf{u}_\beta$ e $\mathbf{u}_\Theta$ denotam as derivadas parciais de primeira ordem da função de verossimilhança relativamente a β e Θ calculadas no ponto ($\beta^{(k-1)}$, $\Theta^{(k-1)}$) e $\mathbf{J}_{\beta\beta}, \mathbf{J}_{\beta\Theta}, \mathbf{J}_{\Theta\beta}$ e $\mathbf{J}_{\Theta\Theta}$ denotam as correspondentes derivadas parciais de segunda ordem calculadas no mesmo ponto. As iterações são interrompidas quando a distância (segundo algum critério) entre as estimativas obtidas em dois passos consecutivos do processo iterativo atinge um valor pré-especificado. No caso do método "Scoring" de Fisher, o procedimento iterativo pode ser simplificado e a expressão (3.2) pode ser utilizada apenas para estimar os parâmetros referentes à estrutura de covariância, Θ, sendo os parâmetros de posição, β, estimados através da expressão (3.1).

Inferências sobre o vetor de parâmteros β podem ser feitas a partir da distribuição (aproximada) do estimador $\tilde{\beta}$, que é Normal multivariada com vetor de médias β e uma matriz de covariância que pode ser estimada por $(\sum_{i=1}^{n} \mathbf{X}_i^t \Sigma_i^{-1}(\tilde{\Theta}) \mathbf{X}_i)^{-1}$ onde $\tilde{\Theta}$ é um estimador consistente de Θ.

Para dados balanceados em relação ao tempo, detalhes sobre os métodos de estimação podem ser encontrados em diversas fontes como Winer (1971), Morrison (1976) ou mais recentemente em Crowder and Hand (1990). Nos casos mais gerais, sugerimos os trabalhos de Jennrich and Schluchter (1986), Laird, Lange and Stram (1987), Andreoni (1989), ou Jones (1993), por exemplo. Procedimentos para o ajuste de modelos estatísticos a dados longitudinais estão incorporados na maioria dos pacotes de "software" estatístico. No BMDP, as subrotinas 2V e 4V podem ser utilizadas para analizar dados balanceados em relação ao tempo; isso também pode ser feito através da subrotina GLM do SAS. No caso geral, a análise pode ser conduzida através da subrotina 5V no BMDP ou MIXED no SAS, por exemplo. Em Andreoni (1989) podem-se encontrar programas fontes para implementação dos algoritmos mencionados para modelos de efeitos aleatórios; esses programas estão escritos na linguagem matricial (CM) do pacote NTIA da EMBRAPA.

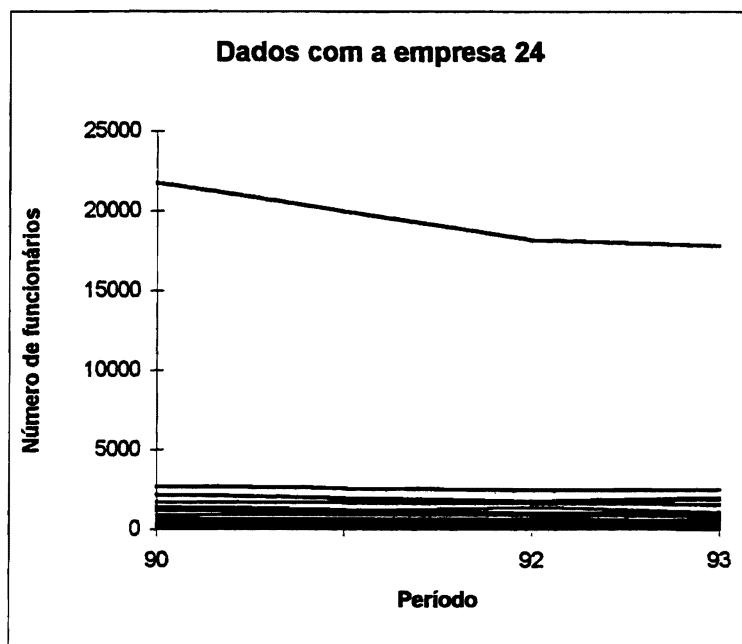
4. Exemplo

Os dados apresentados em anexo correspondem ao número de funcionários empregados em 38 empresas durante os anos de 1990, 1992 e 1993. Eles foram obtidos dos arquivos do IBGE e são utilizados aqui apenas com a finalidade de ilustrar uma análise de dados longitudinais. O objetivo é estudar a evolução do número de funcionários ao longo do período mencionado, assumindo que essas 38 empresas possam ser consideradas como uma amostra aleatória simples de uma população de interesse.

	Número de funcionários no ano de		
Empresa	1990	1992	1993
1	817	1402	1024
2	640	520	482
3	1172	1041	941
4	1444	1016	1059
5	85	23	84
6	596	681	483
7	337	289	229
8	1671	1719	1499
9	462	544	414
10	1263	963	830
11	264	483	549
12	619	542	491
13	633	598	592
14	106	175	161
15	385	442	404
16	189	207	201
17	295	250	243
18	288	247	237
19	533	449	438
20	122	135	112
21	1732	1635	1564
22	658	627	577
23	2716	2469	2450
24	21761	18186	17812
25	2178	1705	1898
26	238	280	105
27	266	294	325
28		417	332
29	2177	1763	1964
30	51	71	14
31	311	371	466
32	136	136	12
33	354	275	238
34	254	272	243
35	349	355	355
36		103	0
37			0
38	922	904	777

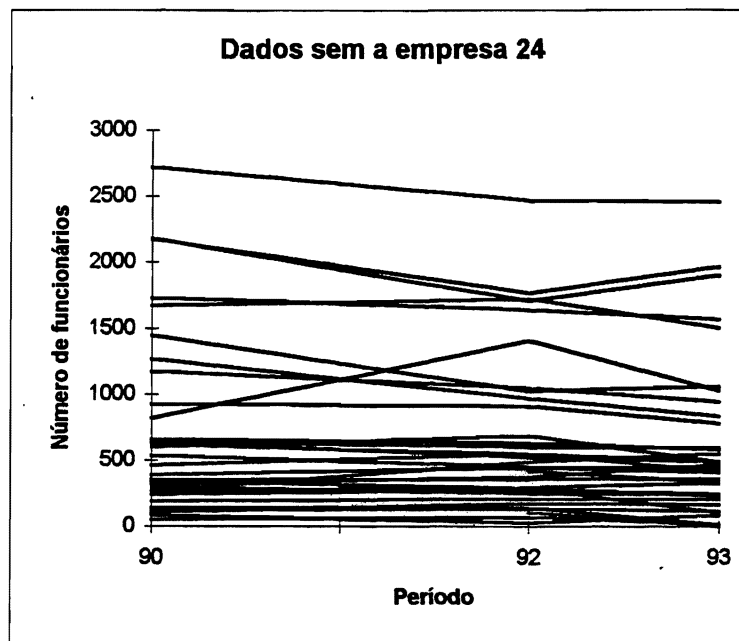
Nas Figuras 4.1 e 4.2 apresentamos os chamados **diagramas paralelos de dispersão** ("parallel plots") com e sem a inclusão de uma das empresas. Esses gráficos podem ser utilizados com a finalidade de identificar possíveis modelos para a estrutura de covariância e/ou de detectar perfis de resposta discrepantes. O leitor poderá consultar Rao and Rao (1966), Weiss and Lazaro (1992) ou Suyama (1995) para detalhes sobre a utilização de

métodos gráficos na identificação de modelos para análise de dados longitudinais. O perfil correspondente à empresa desconsiderada na construção na Figura 4.2 (identificada pelo número 24 no conjunto de dados) pertence a essa categoria; os dados correspondentes foram eliminados do restante da análise.

Figura 4.1: Diagrama paralelo de dispersão para as 38 empresas.

Os números médios de funcionários e correspondentes erros padrões para as 37 empresas consideradas nos anos de 1990, 1992 e 1993 são respectivamente 713 ± 118, 650 ± 98 e 589 ± 98. Pode-se observar uma tendência decrescente das médias nesses três anos. Se desconsiderarmos a possível correlação entre as observações intra-unidades amostrais, a comparação dessas médias via uma ANOVA não evidencia que as diferenças observadas sejam significativas (p=0.7031). A matriz de correlações intra-unidades amostrais é

	1990	**1992**	**1993**
1990	1.000	0.966	0.982
1992	0.966	1.000	0.984
1993	0.982	0.984	1.000

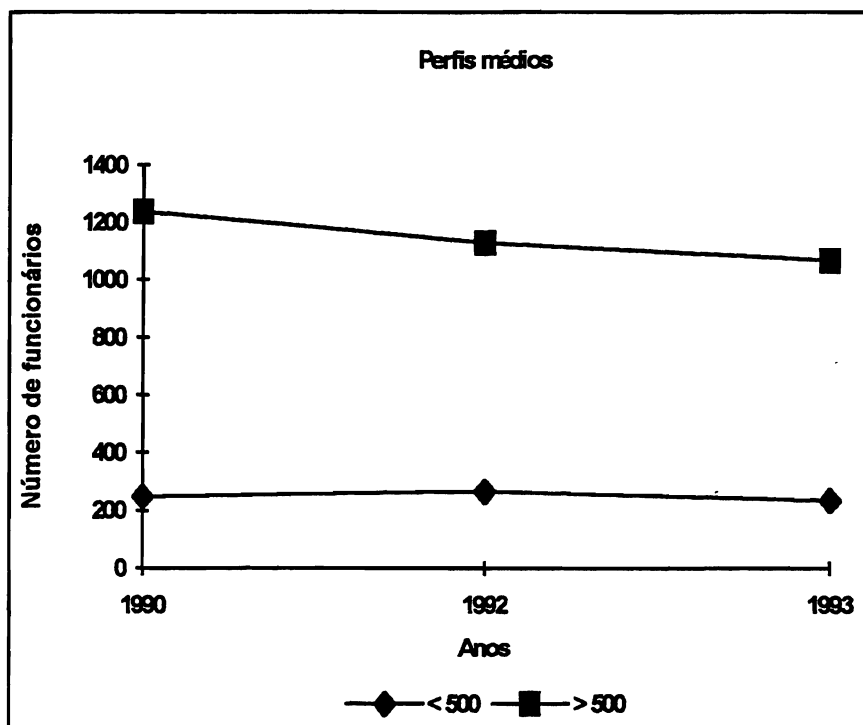
indicando que não só os dados são altamente correlacionados como também que um modelo com estrutura de covariância uniforme poderia ser considerado na análise. Esse modelo também é sugerido pelo padrão de paralelismo dos perfis de resposta da Figura 4.2. Em função dessas indicações comparamos as médias de interesse através de uma ANOVA para medidas repetidas

Figura 4.2: Diagrama paralelo de dispersão excluindo a Empresa de número 24.

utilizando o procedimento GLM do pacote estatístico SAS. Para efeito de ilustração, as matrizes de especificação seriam dadas por $\mathbf{X}_i = \mathbf{I}_3$ e o vetor de parâmetros por $\boldsymbol{\beta} = (m_{90}, m_{92}, m_{93})^t$ onde m_j representa o número médio de funcionários no ano $j, j = 90, 92, 93$. Esse procedimento elimina as empresas com observações incompletas da análise; a inclusão dessas empresas pode ser considerada através de outros procedimentos mais sofisticados; optamos pela alternativa mais simples em função do pequeno número de observações incompletas e do caráter puramente ilustrativo do exemplo. Neste caso, o resultado detecta uma diferença altamente significativa (p=0.0096) entre as médias em questão, sugerindo uma tendência decrescente para o número médio de funcionários empregados.

Uma análise complementar, também com propósitos puramente ilustrativos, envolve a classificação das empresas como **pequenas** ou **grandes** consoante o número de funcionários em 1990 (no exemplo, adotamos arbitrariamente, 500 como ponto de corte). Os números médios de funcionários nesses dois grupos estão indicados abaixo.

Tamanho da empresa	Número médio de funcionários em		
em 1990	1990	1992	1993
Pequena	250	269	244
Grande	1236	1127	1067

Na Figura 4.3 estão representados os chamados **perfis médios**, que essencialmente correspondem aos conjuntos de valores médios associados a cada grupo.

Figura 4.3: Número médio de funcionários

Um exame desses perfis médios indica uma tendência decrescente do número médio de funcionários para as empresas grandes, em contraposição a uma certa estabilidade para as empresas pequenas ao longo do período investigado. Uma Análise de Perfis (ver Winer (1971) ou Singer e Andrade (1986), por exemplo) para esses dados confirma a significância estatística dessa diferença de comportamento (com $p<0.0082$ para o teste da hipótese de paralelismo dos perfis médios, que corresponde à hipótese de inexistência de interação entre Tamanho e Tempo no jargão estatístico) e sugere que o número médio de funcionários das empresas pequenas pode ser considerado estável ($p>0.6560$) com valor 254 entre 1990 e 1993 ao passo que decresce ($p<0.0003$) com uma taxa de 56 por ano para as empresas grandes no mesmo período. A Análise de Perfis é baseada num modelo do tipo "split-plot" (com dois fatores - Tamanho e Tempo) que induz uma estrutura uniforme para a matriz de covariância intra unidades amostrais. Os testes obtidos sob esse modelo continuam válidos mesmo sob condições menos restritivas (i.e. de **esfericidade**) para a estrutura de covariância das observações realizadas nas mesmas unidades amostrais. Além disso, testes aproximados para as hipóteses de interesse podem ser obtidos mediante correções nos graus de liberdade dos testes originais, mesmo sob condições ainda mais gerais. Uma ANOVA usual com os mesmos dois fatores não detecta as diferenças acima.

Agradecimentos

Este trabalho foi parcialmente finaciado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq. O autor agradece a colaboração de Adriana Sañudo na análise dos dados.

Referências Bibliográficas

Andreoni, S. (1989). Modelos de efeitos aleatórios para análise de dados longitudinais não balanceados em relação ao tempo. Dissertação de mestrado. São Paulo: Departamento de Estatística, Instituto de Matemática e Estatística, USP.

Crowder, M.J. and Hand, D.J. (1990). **Analysis of Repeated Measures**. London: Chapman and Hall.

Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994). **Analysis of longitudinal data**. Oxford: Clarendon Press.

Domenech, C.H. (1989). Métodos exploratórios e utilização de medidas resumo para análise de dados longitudinais. Dissertação de mestrado. São Paulo: Departamento de Estatística, Instituto de Matemática e Estatística, USP.

Duncan, G.J. and Kalton, G. (1987). Issues of design and analysis of surveys acroos time. **International Statistical Review, 55**, 97-117.

Goldstein, H. (1979). **The design and analysis of longitudinal studies**. London: Academic Press.

Graybill, F.A. (1976). **Theory and applications of the linear model**. North Scituate, Mass.: Duxbury Press.

Jennrich, R.L. and Schluchter, M.D. (1986). Unbalanced repeated-measures models with structured covariance matrices. **Biometrics, 42**, 805-820.

Jones, R.H. (1993). **Longitudinal Data with Serial Correlation: a State Space Approach**. London: Chapman and Hall.

Kenward, M.G. (1987). A method for comparing profiles of repeated measurements. **Applied Statistics, 36**, 296-308.

Koch, G.G., Singer, J.M. and Stokes, M.E. (1992). Some aspects of weighted least squares analysis for longitudinal categorical data. In **Statistical Models for Longitudinal Studies of Health**, eds. J.H. Dwyer, M. Feinleib, P. Lippert and H. Hoffmeister. (Monographs in Epidemiology and Biostatistics. Vol. 16). New York: Oxford University Press, 215-258.

Laird, N.M., Lange, N. and Stram, D. (1987). Maximum likelihood computations with repeated measures: application of the EM algorithm. **Journal of the American Statistical Association, 82**, 97-105.

Laird, N.M. and Ware, J.H. (1982). Random effects models for longitudinal data. **Biometrics, 38**, 963-974.

Morrison, D.F. (1976). **Multivariate Statistical Methods, 2nd. edition**. New York: McGraw-Hill.

Rao, C.R. (1967). Least squares theory using estimated dispersion matrix and its application to measurement of signals. In **Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability,1,** eds. L.M. LeCam and J. Neyman. 355-372. Berkeley: University of California Press.

Rao, M.N. and Rao, C.R. (1966). Linked cross-sectional study for determining norms and growth-rates - a pilot survey of Indian school-going boys. **Sankhya, B28**, 237-258.

Rochon, J. and Helms, R.W. (1989). Maximum likelihood estimation for incomplete repeated measures experiments under an ARMA covariance structure. **Biometrics, 45**, 207-218.

Singer, J.M. e Andrade, D.F. (1986). **Análise de Dados Longitudinais**. São Paulo: Associação Brasileira de Estatística.

Suyama, E. (1995). Identificação de um modelo de efeitos aleatórios. Tese de doutoramento. São Paulo: Departamento de Estatística, Instituto de Matemática e Estatística, USP.

Weiss, R.E. and Lazaro, C.G. (1992). Residual plots for repeated measures. **Statistics in Medicine, 11**: 115-124.

Winer, B.J. (1971). **Statistical Principles in Experimental Design**, 2nd edition. New York: McGraw Hill.